

Enhancing Privacy of Confidential Data using K Anonymization

R.Vidyabanu¹, Divya Suzanne Thomas² and N.Nagaveni³

¹ Sri Krishna College of Engineering and Technology, Department of Applied Sciences, Coimbatore, India
Email - vidhyabanu@yahoo.com

² Sri Krishna College of Engineering and Technology, Department of Applied Sciences, Coimbatore, India

³ Coimbatore Institute of Technology, Department of Mathematics, Coimbatore, Tamilnadu, India

Abstract—Recent advances in the field of data collection and related technologies have inaugurated a new era of research where existing data mining algorithms should be reconsidered from a different point of view, this of privacy preservation. Much research has been done recently on privacy preserving data mining (PPDM) based on perturbation, randomization and secure multiparty computations and more recently on anonymity including k-anonymity and l-diversity.

We use the technique of k-Anonymization to de-associate sensitive attributes from the corresponding identifiers. This is done by anonymizing the linking attributes so that at least k released records match each value combination of the linking attributes. This paper proposes a k-Anonymization solution for classification. The proposed method has been implemented and evaluated using UCI repository datasets.

After the k-anonymization solution is determined for the original data, classification, a data mining technique using the ID3 algorithm, is applied on both the original table and the compressed table. The accuracy of the both is compared by determining the entropy and the information gain values. Experiments show that the quality of classification can be preserved even for highly restrictive anonymity requirements.

Index Terms—k-Anonymization, privacy, masking, classification

I INTRODUCTION

In today's information society, given the unprecedented ease of finding and accessing information, protection of privacy has become a very important concern. In particular, large databases that include sensitive information (e.g., health information) have often been available to public access, frequently with identifiers stripped in an attempt to protect privacy. However, if such information can be associated with the corresponding people's identifiers, perhaps using other publicly available databases, then privacy can be seriously violated.

The notion of k-anonymity was first proposed in [2]. In general, a cost metric is used to measure the data distortion of anonymization. Two types of cost metric have been considered. The first type, based on the notion of minimal generalization [3] [4], is independent of the purpose of the data release. The second type factors in the purpose of the data release such as

classification [5]. The goal is to find the optimal k-anonymization that minimizes this cost metric. In general, achieving optimal k-anonymization is NP-hard [6] [7]. Greedy methods were proposed in [8] [9] [10] [11]. Scalable algorithms (with the exponential complexity the worst-case) for finding the optimal k-anonymization were studied in [3] [4] [5].

The optimal k-anonymization is not suitable to classification where masking structures and masking noises have different effects: the former seems to damage classification whereas the latter helps classification. It is well known in data mining and machine learning that the unmodified data, which has the lowest possible cost according to any cost metric, often has a worse classification than some generalized (i.e., masked) data. Similarly, less masked data could have a worse classification than some more masked data. The optimal k-anonymization seeks to minimize the error on the training data, thus over-fits the data, subject to the privacy constraint. Neither the over-fitting nor the privacy constraint is relevant to the classification goal that seeks to minimize the error on future data.

II PRELIMINARIES

A. Masking

To transform the table T to satisfy the anonymity requirement, we consider three types of masking operations on the attributes D(j) in U[QID] where QID refers to the quasi identifier.

Generalize

D(j) is a categorical attribute with a taxonomy tree. A leaf node represents a domain value and a parent node represents a less specific value. A generalized D(j) can be viewed as a cut through its taxonomy tree. A cut of a tree is a subset of values in the tree that contains exactly one value on each root-to leaf path. This type of generalization does not suffer from the interpretation difficulty discussed earlier.

Suppress:

D(j) is a categorical attribute with no taxonomy tree. The suppression of a value on D(j) means replacing all occurrences of the value with the special value. All suppressed values on D(j) are represented by the same,

which is treated as a new value in $D(j)$ by a classification algorithm.

Discretize:

$D(j)$ is a continuous attribute. The discretization of a value v on $D(j)$ means replacing all occurrences of v with an interval containing the value.

B.ID3 Algorithm

Each non-leaf node of a decision tree corresponds to an input attribute, and each arc to a possible value of that attribute. A leaf node corresponds to the expected value of the output attribute when the input attributes are described by the path from the root node to that leaf node.

In a “good” decision tree, each non-leaf node should correspond to the input attribute which is the most informative about the output attribute amongst all the input attributes not yet considered in the path from the root node to that node. This is because we would like to predict the output attribute using the smallest possible number of questions on average.

Entropy is used to determine how informative a particular input attribute is about the output attribute for a subset of the training data. Entropy is a measure of uncertainty in communication systems. It is fundamental in modern information theory.

III PROPOSED K-ANONYMIZATION TECHNIQUE

A. Masking operation

To transform T to satisfy the anonymity requirement, we consider three types of masking operations on the attributes $D(j)$ in $U[QID]$ is done as mentioned above.

B. Anonymisation by top-down refinement

For given Data Set, Consider the ‘p’ Quasi identifiers $QID(1), \dots, QID(p)$ on table T . $A(qid(i))$ denotes the number of data records in T that share the value $qid(i)$ on $QID(i)$. The anonymity of $QID(i)$, denoted $A(QID(i))$, is the smallest $a(qid(i))$ for any value $qid(i)$ on $QID(i)$. A table T satisfies the anonymity requirement $(QID(1), k(1), \dots, QID(p), k(p))$ if $A(QID(i)) \geq k(i)$ where $1 \leq i \leq p$ and where $k(i)$ is the anonymity threshold on $QID(i)$ specified by the data provider. We make note of the no. of the records or attribute values for each quasi identifiers that are in the given database and the check on the anonymity requirement Threshold level $k(i)$ for each $QID(i)$.

C. Refinement of the masked datasets

A table T can be masked by a sequence of refinements starting from the most masked state in which each attribute is either generalized to the top most value, or suppressed to the special value, or represented by a single interval. This method iteratively refines a masked value selected from the current set of cuts, suppressed values and intervals, until violating the anonymity requirement. Each refinement increases the

information and decreases the anonymity since records with specific values are more distinguishable. The key is selecting the best refinement at each step with both impacts considered.

The notion of refinement on different types of attributes $D(j)$ that belongs to $UQID(i)$ and defining selection criterion for a single refinement is formally defined below.

Refinement for Generalization: Consider a categorical attribute $D(j)$ with a user-specified taxonomy tree. Let $child(v)$ be the set of child values of v in a user-specified taxonomy tree. A refinement, written $v \rightarrow child(v)$, replaces the parent value v with the child value in $child(v)$ that generalizes the domain value in each (generalized) record that contains v .

Refinement for Suppression: For a categorical attribute $D(j)$ without taxonomy tree, a refinement, refers to disclosing one value v from the set of suppressed values $Sup(j)$. Let R_j denote the set of suppressed records that currently contain $per(j)$. Disclosing v means replacing $per(j)$ with v in all records in $R(j)$ that originally contain v .

Refinement for Discretization: For a continuous attribute, refinement is similar to that for generalization except that no prior taxonomy tree is given and the taxonomy tree has to be grown dynamically in the process of refinement. Initially, the interval that covers the full range of the attribute forms the root. The refinement on an interval v , written for all $child(v)$, refers to the optimal split of v into two child intervals $child(v)$ that maximizes the information gain. The anonymity is not used for finding a split good for classification. This is similar to defining a taxonomy tree where the main consideration is how the taxonomy best describes the application. Due to this extra step of identifying the optimal split of the parent interval, we treat continuous attributes separately from categorical attributes with taxonomy trees. A refinement is valid (with respect to T) if T satisfies the anonymity requirement after the refinement. A refinement is beneficial (with respect to T) if more than one class is involved in the refined records. A refinement is performed only if it is both valid and beneficial. Therefore, a refinement guarantees that every newly generated qid has a $a(qid)$, k .

D. Computation of infogain, anyloss and score of the refinement:

InfoGain(v): defined as

$$InfoGain(v) = I(R_v) - \sum_c \frac{|R_c|}{|R_v|} I(R_c) \quad (1)$$

where $I(R_x)$ is the entropy of $I(R_x) =$

$$\forall_{cls} \frac{freq(R_x, cls)}{|R_x|} * |\log \frac{freq(R_x, cls)}{|R_x|}| \quad (2)$$

$freq(R_x; cls)$ is the number of data records in R_x having the class cls . Intuitively, $I(R_x)$ measures the entropy (or impurity) of classes in R_x . The more dominating the majority class in R_x is, the smaller $I(R_x)$ is (i.e., less entropy in R_x). Therefore, $I(R_x)$ measures the error because non-majority

AnonyLoss(v): defined as

$$AnonyLoss(v) = \frac{avg\{A(QID_j) - A_v(QID_j)\}}{|QID_j|} \quad (3)$$

where $A(QID_j)$ and $A_v(QID_j)$ represent the anonymity before and after refining v . $avg\{A(QID_j) - A_v(QID_j)\}$ is the average loss of anonymity for all QID_j that contain the attribute of refinement process to heuristically maximize the classification goal. Consider a refinement for all child(v) where for all D_j , D_j is a categorical attribute with a user-specified taxonomy tree or D_j is a continuous attribute with a dynamically grown taxonomy tree.

The refinement has two effects: it increases the information of the refined records with respect to classification, and it decreases the anonymity of the refined records with respect to privacy. These effects are measured by information gain denoted by $InfoGain(v)$ in (1), and anonymity loss denoted by $AnonyLoss(v)$ in (3). v is a good candidate for refinement. If $InfoGain(v)$ is large and $AnonyLoss(v)$ is small. Our selection criterion is choosing the candidate v , for the next refinement, that has the maximum information-gain/ anonymity loss trade-off, defined as

$$Score(v) = InfoGain(v) / AnonyLoss(v) + 1 \quad (4)$$

1 is added to $AnonyLoss(v)$ to avoid division by zero. Each choice of $InfoGain(v)$ and $AnonyLoss(v)$ gives a trade-off between classification and anonymization. It should be noted that $Score$ is not a goodness metric of k -anonymization. Infact, it is difficult to have a closed form metric to capture the classification goal on future data.

E. Implementation of ID3 algorithm

ID3 algorithm is used to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a given sets, we introduce a metric information gain.

To find an optimal way to classify a learning set, what we need to do is to minimize the questions asked. Thus, we need some function which can measure which questions provide the most balanced splitting.

IV RESULTS

Table 1 depicts the runtime of top down refinement using generalization, suppression and discretization for 200K to 1M data records and based on two types of anonymity requirements. AllAttQID refers to the single quasi identifiers having all 14 attributes. This is one of the most time consuming settings because of the largest number of candidate refinements to consider at each iteration. Top down refinement requires more iterations to reach a solution, hence more runtime. Top down refinement takes approximately 80 seconds to transform 1M records. Compared to AllAttQID, top down refinement becomes less efficient for handling multiple quasi identifiers. The implementation results summarized below. Fig 1 shows the time taken for masking when executed with different number of records.

Data Source: UCI Repository: Adult dataset
Fields: Age, Work Class, Sex
Operation: Masking
Suppression: Work Class, Sex
Discretization: Age

TABLE 1 TIME TAKEN IN SECONDS

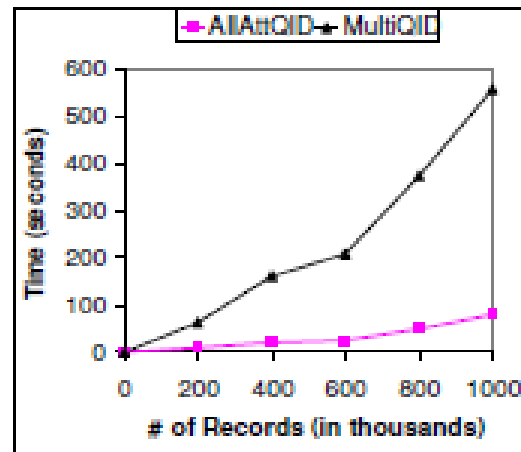


Figure 1 . for masking

Table two summarizes the final results on applying the ID3 algorithm on the compressed and the original data for a series of inputs. A comparison of the accuracy of classification of transformed data with the original data is depicted in fig 2.

Data Source: UCI Repository: Adult dataset
Fields: Age, Work Class, Sex
Operation: Masking
Suppression: Work Class, Sex
Discretization: Age
Generalization: Education
Operation: ID3 Algorithm
Accuracy Calculation:
Accuracy % = No. of records selected / Entropy

Table 2

Decision Tree Accuracy

No. of records	Original Table	Compressed Table
10		
50		
100		
150	.33	.66
200	.55	
250	.66	.80
300	.98.15	.66
350	.97.62	.59.97

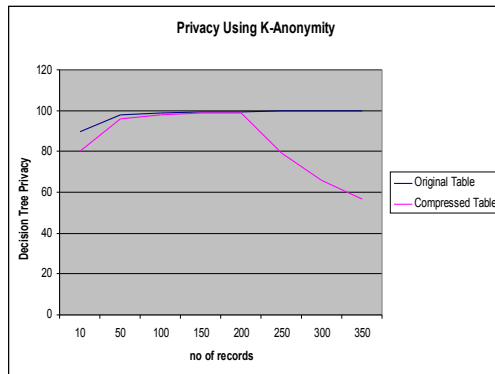


Figure 2 . Accuracy using k-Anonymity

V CONCLUSION

The problem of ensuring individual's anonymity while releasing person-specific data for classification analysis is considered. The previous optimal k-anonymization based on a closed form of cost metric does not address this classification requirement. The proposed method can be used to hide valuable information while presenting it on a publicly accessible place like internet. The results shows that when classification is applied to both the original table and transformed table, the accuracy level is not too low for the transformed table when compared to that of the original table. The proposed privacy preserving transformation preserved the nature of the data even in the transformed form. The classification accuracy while using the transformed data is almost equal to that of the original dataset.

REFERENCES

- 1]. Benjamin C. M. Fung, Ke Wang, and Philip S. Yu, Fellow "Anonymizing Classification Data for Privacy Preservation" *IEEE transactions on knowledge and engineering* 2007.
- 2]. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information" *Proc. of the 17th ACM SIGACTSIGMOD- SIGART Symposium on Principles of Database Systems (PODS 98)*, Seattle, WA, June 1998, p. 188.
- 3]. P. Samarati, "Protecting respondents' identities in microdata release", *IEEE Transactions on Knowledge Engineering (TKDE)*, vol. 13, no. 6, 2001, pp. 1010.1027
- 4]. L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression". *International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems*, vol. 10, no. 5, pp. 571.588, 2002.
- 5]. R. J. Bayardo and R. Agrawal, "Data privacy through optimal kanonymization" *Proc. of the 21st International Conference on Data Engineering (ICDE)*, Tokyo, Japan, April 2005, pp. 217.228.
- 6]. G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, "Approximation algorithms for k-anonymity", *Journal of Privacy Technology*, no. 20051120001, November 2005
- 7]. A. Meyerson and R. Williams, "On the complexity of optimal kanonymity". *Proc. of the 23rd ACM Symposium on Principles of Database Systems (PODS)*, 2004, pp. 223.228.
- 8]. L. Sweeney, "Data_y" A system for providing anonymity in medical data," *Proc. of the International Conference on Database Security*, 1998, pp. 356.381.
- 9]. V. S. Iyengar, "Transforming data to satisfy privacy constraints", *Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, AB, Canada, July 2002, pp. 279.288.
- [10]. K. Wang, P. Yu, and S. Chakraborty, "Bottom-up generalization: a data mining solution to privacy protection", *Proc. of the 4th IEEE International Conference on Data Mining (ICDM)*, November 2004.
- [11]. B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation". *Proc. of the 21st International Conference on Data Engineering (ICDE)*, Tokyo, Japan, April 2005, pp. 205.216.